



Heriot-Watt University
Research Gateway

It's Good to Chat? Evaluation and Design Guidelines for Combining Open-Domain Social Conversation with Task-Based Dialogue in Intelligent Buildings

Citation for published version:

Gunson, N, Sieinska, W, Walsh, C, Dondrup, C & Lemon, O 2020, It's Good to Chat? Evaluation and Design Guidelines for Combining Open-Domain Social Conversation with Task-Based Dialogue in Intelligent Buildings. in *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents.*, 24, Association for Computing Machinery, New York, NY, USA, 20th ACM International Conference on Intelligent Virtual Agents 2020, Glasgow, United Kingdom, 20/10/20.
<https://doi.org/10.1145/3383652.3423889>

Digital Object Identifier (DOI):

[10.1145/3383652.3423889](https://doi.org/10.1145/3383652.3423889)

Link:

[Link to publication record in Heriot-Watt Research Portal](#)

Document Version:

Peer reviewed version

Published In:

IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents

Publisher Rights Statement:

Copyright © 2020 ACM

General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

It's Good to Chat? Evaluation and Design Guidelines for Combining Open-Domain Social Conversation with Task-Based Dialogue in Intelligent Buildings

Nancie Gunson
Heriot-Watt University
Edinburgh, U.K.
n.gunson@hw.ac.uk

Weronika Sieińska
Heriot-Watt University
Edinburgh, U.K.
w.sieinska@hw.ac.uk

Christopher Walsh
Heriot-Watt University
Edinburgh, U.K.
c.walsh.1@research.gla.ac.uk

Christian Dondrup
Heriot-Watt University
Edinburgh, U.K.
c.dondrup@hw.ac.uk

Oliver Lemon
Heriot-Watt University
Edinburgh, U.K.
o.lemon@hw.ac.uk

ABSTRACT

We present and evaluate a deployed conversational AI system that acts as a host of a working public building on a university campus. The system combines open-domain social chat with task-based conversation regarding navigation in the building, live resource updates (e.g. available computers), and events in the building. We investigated the impact of open-domain social chat on task completion and user preferences by comparing the combined system with a task-only version. We find that there is no significant difference in task completion or several aspects of user preference between the two systems, but that users would be significantly happier to talk to the task-only system in the future. This suggests that the “walk-up” public setting and workplace nature of the environment creates a markedly different use case to the in-home, and more individual and private “companion/assistant” setting which is commonly assumed for systems like Alexa. We discuss the implications for the design of conversational systems in other public settings.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; Empirical studies in HCI; • **Computing methodologies** → *Dis-course, dialogue and pragmatics*.

KEYWORDS

conversational artificial intelligence, intelligent buildings, smart speakers, empirical studies, voice assistants, chatbots

ACM Reference Format:

Nancie Gunson, Weronika Sieińska, Christopher Walsh, Christian Dondrup, and Oliver Lemon. 2020. It's Good to Chat? Evaluation and Design Guidelines for Combining Open-Domain Social Conversation with Task-Based

Dialogue in Intelligent Buildings. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk*. ACM, New York, NY, USA, 8 pages. <https://doi.org/XX.XXXX/YYYYYYY.YYYYYY>

1 INTRODUCTION

Conversational Agents or Spoken Dialogue Systems have been a topic of research for decades [7, 11, 25]. Much of the research has focused on task-oriented systems, where the emphasis is on completing a user's goal (such as flight booking) through a series of dialogue turns [19, 27]. Conversations are designed to be short and functional in order to support completion of the task, with little or no attempt made to entertain or establish a relationship with the user.

Chatbots, on the other hand, (e.g. ALICE [1]) are designed to promote extended, unstructured conversation more characteristic of human-human interaction, often with no particular ‘goal’ or task beyond entertaining and engaging the user. Significant strides have been made in this area in recent years, in part due to initiatives such as the Amazon Alexa Socialbot Grand Challenge [10].

Largely, however, conversational systems remain divided between those classed as task-oriented, and chatbots focused solely on entertainment [4]. Given that human conversations usually interleave social content with task content [20] this division is artificial and unnatural. To create more engaging interactions, it seems apparent that conversational systems must support both task completion and entertaining chat, and allow the user to switch effortlessly between the two. However, there is surprisingly limited empirical evidence on the subject.

Voice-based intelligent personal assistants released by major tech companies such as Amazon's Alexa, Google's Assistant, and Samsung's Bixby provide access to both entertainment and task-based interaction, but these features have to be requested or invoked separately and are often limited to tasks or jokes etc. that are completed with a single user-system turn. These systems are predominately used within the home and are targeted at a specific known user.

Here, the domain of interest is a public building, specifically a smart building on a university campus, where many different people can walk up and use the system. We present and evaluate a conversational agent deployed on a Google Home Mini that acts

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN AAA-B-CCCC-CCCC-X/YY/ZZ...\$XX.XX

<https://doi.org/XX.XXXX/YYYYYYY.YYYYYY>

as host of the building, and has the ability to combine state-of-the-art open-domain social conversation with task-based assistance relevant to the context. Such a setting can provide us with valuable insights for design in similar situations, such as communal spaces in other public or professional buildings, and for example elder-care facilities¹.

In an empirical evaluation with 32 users we investigate the benefits of this combined approach by comparing our socially-enabled bot to a task-only version of the system. The effect on the user experience is measured using a range of subjective and objective measures, including user attitudes towards the two different systems, their explicit preference between the two and measures of conversational efficiency and dialogue quality.

2 RELATED WORK

Some limited work exists on combining chat and task-based dialogue. In [26], for example, it was found that interleaving task and social content tends to increase both users' engagement and task-performance. Here the 'task' was audience research relating to movies; establishing users' preferences and recent viewing history in order to provide them with a movie recommendation, and the film distributors with useful audience data. The task itself was arguably therefore fairly social in nature. All interactions were text-based and the user evaluation was carried out by Amazon Mechanical Turk workers in a home setting.

Previous research in the context of intelligent buildings showed that the inclusion of social dialogue had a positive effect on users' compliance with suggestions from an Embodied Conversational Agent (ECA) for energy-efficiency behaviours [9]. Here the context was an office environment and social dialogue was limited to two turns of small-talk at the start of the interaction (*"Hi, how are you?"* and *"What's your name?"*).

Research on relational agents has investigated the use of small talk in "greasing the wheels" of task-based dialogue with a real-estate ECA (called 'REA') and found positive effects on engagement and trust [2]. Similarly, an evaluation of the museum guide 'Tinker' showed that when the ECA employed relational behaviours that included social dialogue this led to significantly greater engagement by museum visitors [3]. Here, input was via multiple choice utterance input screens. Neither system had the capacity for open-domain chat.

Pecune et al. [17] carried out a field trial of a socially aware ECA who acted as host at an international conference. The system is organised around a task-based pipeline and a social one. However, on closer inspection the social aspects are limited to 'decorating' the task content in terms of the style of the sentence, rather than the content itself, and during the trial itself the system was only partially automated.

In one of the most relevant examples [16] researchers investigated combining task and social dialogue in a robot intended for deployment as a guide in a shopping mall. Their evaluation showed that the hybrid task+chat system was rated significantly more pleasant and better met the users' expectations than a task-only version. This is encouraging given that, in the words of the authors, the robot had "rather simple" chat capabilities. As with much of the

other research on this topic, also, it was carried out in the context of a multimodal interface.

The conversational agent examined in this paper is currently deployed as a voice-only assistant, on Google Home Mini devices which in a real-world deployment have the advantage of being widely-available, low-cost devices.

Voice-only assistants in the form of smart speakers have been much studied in the home setting [18, 21]; however, there is little evidence of research on their use in public spaces. In one study, Lopatovska and Oropeza [13] investigated the adoption of Amazon's Alexa by students and staff in the foyer of an academic building. They found that only 10% of foot traffic interacted with the Echo device, with survey participants reporting lack of need, lack of awareness, and the perceived awkwardness of interacting in a public space as their reasons. Enquiry patterns also differed to those reported by studies in the home with user requests in public consisting largely of questions about Alexa's features, "opinions", and mindless questions. It should be noted, however, that only native Alexa content was available, with no tailored building or university-specific information – or a wider ability to chat.

We present a building host that combines task-based assistance relevant to the context with a level of open-domain chat capability not found in any of the above systems. The system is based on the social bot Alana, which was twice finalist in the Amazon Alexa challenge, and can converse on topics ranging from the latest news headlines to philosophy [5, 14, 15] and has previously proven successful as a foundation for other conversational AI projects [6]. We contribute to the field with an empirical evaluation of the benefits of combining such open-domain chat capability with task-based dialogue in the context of a real-world deployment in a public setting.

3 SYSTEM DESIGN

The GRID building at the Heriot-Watt University campus in Edinburgh is a large, multi-purpose teaching and collaborative space with multiple types of users, predominately students but also staff and visitors from industry. To create our intelligent conversational host for this building we extended the social bot Alana to include several new task-based functionalities specific to the building and its activities [23]. The new features include providing users with information on:

- the GRID building itself (e.g. facilities, rooms, construction date, opening times),
- the location of rooms and directions to them,
- events happening in the building,
- computers available for use – updated live.

Figure 1 presents the architecture of the system. The Alana system is an ensemble of data-driven and rule-based chat bots that compete in parallel to generate a reply to the user's utterance. Retrieval bots draw on a wide range of information sources to produce their potential replies including Wikipedia, Reddit, and a variety of news feeds on NewsAPI. Rule-based bots are used to respond in a controlled, consistent way to specific user queries e.g. in Persona expressing the views, likes, and dislikes of the virtual personality 'Alana'. The Coherence Bot is responsible for maintaining the flow of the conversation, for example by returning to previous topics

¹See the SPRING project: <https://spring-h2020.eu/>

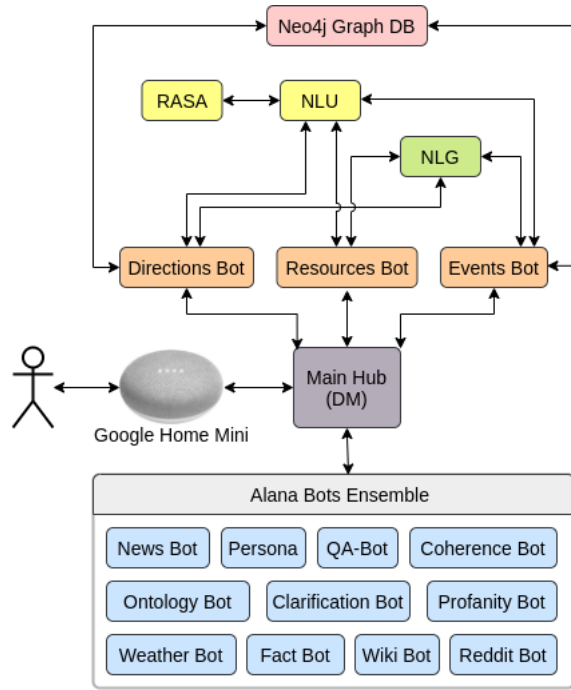


Figure 1: System architecture.

of interest. It also provides conversational ‘drivers’ to progress the conversation after relevant information has been delivered by other bots. Additional bots in the Alana ensemble include those that deal with profanities, handle entity clarifications, etc.

In the GRID project, three new bots have been implemented: Directions, Resources, and Events Bots. The Directions Bot handles requests for locations of rooms and directions to them. It connects to a Neo4j graph database (see Figure 2) from where it retrieves the shortest path to an identified location. The Events Bot handles enquiries about events taking place in the building, either in a specific named room or the building in general. This event information is also stored in the Neo4j database. It is retrieved from a university timetable system and updated daily. The role of the Resources Bot is to provide the user with information about how many computers are currently free in the GRID building’s labs. The data is scraped live from a university website. General, fixed information about the GRID building is encoded in a modified version of the Alana system’s Persona Bot, implemented as AIML templates. It is this bot that responds to e.g. questions about the building’s purpose and facilities.

Following a response to a building-related enquiry, a conversational driver is uttered that ‘advertises’ both the system’s building-specific and other conversational capabilities e.g. “*Did you know that....*”. Users engaged in task dialogue are thus made aware that the option to chat is available, but the decision on whether or not to do so lies with them (user-initiative). In cases where the Coherence bot responds because there has been some problem in the dialogue, it takes the initiative, asking a question or making a statement on the current topic or a new one in order to move the conversation forward. In the Alana-GRID system the list of topics has been expanded to include the building-specific capabilities of the system.

ASR/TTS. Automatic Speech Recognition in the GRID project is handled using the Google Speech API. The system’s responses are voiced using Google Assistant’s female voice 2. ASR errors which mean that a user’s intent cannot be extracted at the NLU stage are handled by the Coherence bot. Where the ASR confidence is very low, users receive a message prompting them to repeat their input e.g. *I’m afraid I didn’t get that. Could you repeat that please?*

NLU. In the Alana system, users’ utterances are parsed using a complex Natural Language Understanding pipeline, described in detail in [5]. Stages include Named Entity Recognition, Noun Phrase extraction, co-reference and ellipsis resolution, and both neural and regex-based intent recognition. In the GRID project, an additional NLU module has been implemented for building-specific Directions, Events and Resources enquiries which is based on the RASA framework². In the Persona Bot we use AIML patterns.

NLG. The Natural Language Generation strategy depends on the specific bot. Information retrieval bots such as News Bot employ summarisation techniques [5] whilst the GRID bots’ responses use robust template-based methods.

Dialogue Management. At every dialogue turn each of the bots attempts to generate a response. The decision as to which response is selected for verbalisation to the user is then handled by the Dialogue Manager. The current selection strategy is based on a bot priority list, although it can also be learned from data [22]. If multiple bots produce a response, the response from the bot with the highest priority is selected. In the Alana-GRID system, the building-specific Directions, Events, Resources and Persona Bots are placed at the top of the list, in that order. The remaining bots follow according to the order of priority employed in [5].

The Neo4j Graph Database. Figure 2 presents the graph representation of how the rooms are connected with each other in the GRID building. To port the system to different buildings, this representation can be changed to reflect the new locations and their spatial connectivity.

4 EVALUATION

4.1 Experiment Conditions

In order to investigate the value of combining open-domain social chat and task-based dialogue in the smart building context, two conditions were compared; the Alana-GRID system described in Section 3, which combines building-specific task-based dialogue with open-domain conversation (task+chat) and a solely task-based system (task-only) deprived of any open-domain conversational ability.

In the task-only condition the bot ensemble described in Figure 1 was restricted to the bots specific to the GRID building (Resources, Events, and Directions Bots, modified Persona Bot, and modified Coherence Bot), together with the Profanity Bot. Here, Coherence Bot mentioned *only* the GRID-specific capabilities of the system.

Example Dialogue 1 illustrates the task+chat condition and includes examples of responses generated by the GRID Persona and Coherence Bots. Dialogue 2 demonstrates the task-only condition

²See <https://rasa.com/>

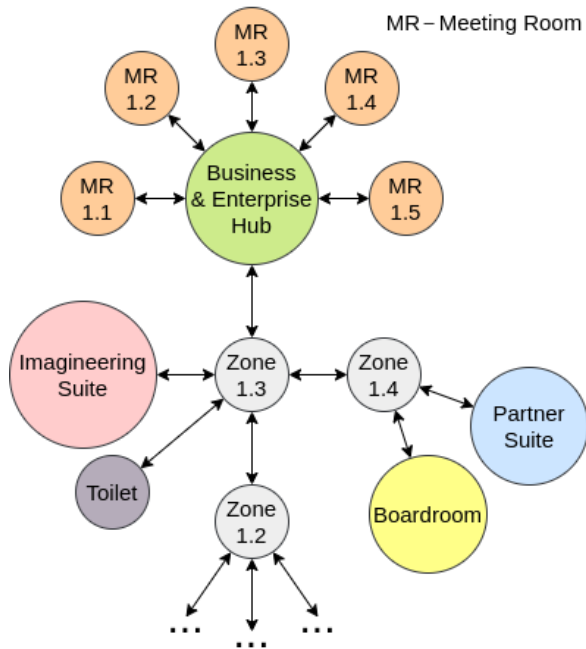


Figure 2: Graph representation of how rooms are connected with each other in the GRID building.

and shows example responses produced by the Directions, Events, and Resources Bots.

4.2 Experiment Procedure

All participants experienced both conditions (task-only and task+chat) in a repeated-measures design. The order of presentation of the two versions was balanced across the group to control for any order effects. With each, participants were asked to carry out the same set of four tasks. These were designed to be realistic and to showcase the main functionalities of the system (enquiries about the Building, Events, Directions and Resources). The tasks were divided into two pairs, so that participants experienced each condition twice. This had the benefit of measuring participants' experience of the two different conditions on repeat as well as first use. The order of the task-pairs, and of the tasks within each pair was balanced across the group. In order to avoid 'putting words into people's mouths' the tasks were inspired with graphical prompts, as illustrated in Figure 3.



Figure 3: Example task prompt.

All interactions were in English. The experiment took place in a meeting room within the GRID building itself. A researcher was present in the room throughout and followed a standardised procedure and script to ensure a consistent experience across participants.

A: Welcome to the GRID building! How can I help you today?
 U: What time does this building close?
 A: (persona) This building never closes. It is open 24/7. Did you know that you can have a chat with me about music? Also, if you're looking for a specific room, you can ask me how to get there.
 U: Actually, I really need a cup of coffee. Do you know where I can get one?
 A: (persona) You can have a cup of coffee at the cafe on the ground floor. It looks like a tuk tuk. In case you didn't know, you can ask me about events happening in this building, current news headlines, or anything else you want.
 U: What is a tuk tuk?
 A: (wiki_bot) I think that an auto rickshaw is a motorized version of the pulled rickshaw or cycle rickshaw. Most have three wheels and do not tilt. They are known by many terms in various countries including auto, baby taxi, bajaj, chand gari, lapa, tuk-tuk, 3wheel or tukxi.
 U: OK, interesting.
 A: (coherence_bot) Do you know that you can ask me about events happening in this building, current news headlines, or anything else you want.
 U: What music do you like?
 A: (persona) I enjoy a lot of music but I like electricity buzzing sounds the most. I love music. Who is a musician that you like?
 U: I like Bob Dylan a lot
 A: (news_bot) A friend of mine told me about this news headline from MTV News about Bob Dylan. Timothée Chalamet will play Bob Dylan in his next Oscar worthy role. Do you want to hear some more about this?
 U: Who is Timothée Chalamet?
 A: (wiki_bot) As far as I know, Timothée Hal Chalamet is an American actor.
 U: What movies do you like?
 A: (persona) My favourite movie is Star Wars. I love the two androids in that film. Which movie do you like?
 U: The new Star Trek movies are pretty good.
 A: (news_bot) Related to Star Trek, I heard this news headline from a friend, 'Star Trek: Picard': Everything you need to watch before the new CBS series. Do you want to learn more?
 U: No, thanks. I've gotta go. Bye!
 A: It was lovely talking to you! Goodbye!

Dialogue 1: Task+chat Condition (U–User, A–Alana system).

Prior to talking to the system participants were informed that the researcher was not allowed to intervene during the interaction, and were encouraged to behave with the system as they would choose to in real life.

Following each interaction, participants completed a short attitude questionnaire, discussed in Section 4.3. At the end of the session they completed an exit questionnaire to determine their preference between versions and gather qualitative information on their experiences, together with demographic data. Sessions lasted a maximum of 30 minutes.

4.3 Dependent Variables

We collected a range of subjective and objective measures. *User Attitude* was assessed using responses to a Likert questionnaire [12] completed by participants after each interaction. The questionnaire consisted of five proposal statements concerning users' happiness

A: *Hello, and welcome to the GRID building! What can I assist you with?*
 U: Hi! What can you do?
 A: (persona) *If you're looking for a room, I can tell you where it is. I can also tell you what events are happening in this building and where you can find a free computer.*
 U: Are there any available computers in the Flex Lab?
 A: (resources_bot) *Sorry, there is Flex Lab 1 and Flex Lab 2 in this building. Which one do you mean?*
 U: Number 2.
 A: (resources_bot) *I'm afraid there are no free computers in the Flex Lab 2 at the moment, but I know that the room Digital Lab has the most free computers, it has 97 PCs available.*
 U: Where is the Digital Lab?
 A: (directions_bot) *The Digital Lab is on the first floor. Would you like directions?*
 U: Yes, please.
 A: (directions_bot) *From the entrance, head across the atrium to the right and take the accessible lift to the first floor, then turn right, the door to the Digital Lab will be in front of you.*
 U: OK, thanks.
 A: (coherence_bot) *I hope you remember that I can provide you with information about events happening in this building.*
 U: Ok, is there anything on today?
 A: (events_bot) *There are 2 events taking place in the Boardroom today. An event called Negotiate Training is scheduled from 10:30 until 12:30. Another event called Virtual Reality Workshop is taking place from 13:15 until 19:15. Is there anything else I can help you with?*
 U: No, I'm good. Thanks, goodbye!
 A: *It was lovely talking to you! Goodbye!*

Dialogue 2: Task-only Condition (U–User, A–Alana system).

to talk to the system again, the degree to which they enjoyed talking to the system, its *usefulness*, perceived *intelligence*, and their level of *trust* in the information provided by the system. We also considered another key measure of overall attitude: participants' explicit preference between the two versions of the system, which was obtained as part of the exit interview.

We considered two categories of objective measures based on the PARADISE framework for evaluation of conversational systems [24]. *Conversational Efficiency* was assessed based on the number of turns and the dialogue duration. *Dialogue Quality* was evaluated using task completion rates as recorded in the logfiles. Tasks were judged complete when the system uttered the relevant information.

4.4 Experiment Hypotheses

Based on previous research, we anticipated the following:

- **H1.** Users will exhibit a more positive attitude towards the system that combines social chat with task-based dialogue, as determined by the user questionnaire.
- **H2.** Correspondingly, the system that combines chat with task-based dialogue will be preferred overall.
- **H3.** Task completion rates will be unaffected by combining social chat with task-based dialogue in the system.
- **H4.** Combining social chat with task-based dialogue will lead to a reduction in conversational efficiency, indicated by an increase in the number of turns and the dialogue duration.

4.5 Participants

A total of 32 students (21 male, 11 female) agreed to take part in the study, ranging in age from 18 years to 27 years (mean 21.21). The vast majority were recruited in person in the GRID building, ensuring they were potential real-life users of the system. 22 were native English speakers. The majority (20) rarely or never use a voice assistant. Each was given a £10 Amazon voucher as reward.

5 RESULTS

5.1 Quantitative Data

5.1.1 User Attitude. In general, participants gave positive scores to both versions (see Figure 4). However, there was little indication of a consistent difference between the two, contradicting our first hypothesis H1. There were a total of 212 positive ratings (those at 4 or 5 on the Likert scale) for the task-only condition compared to 214 for the task+chat version; an almost identical number.

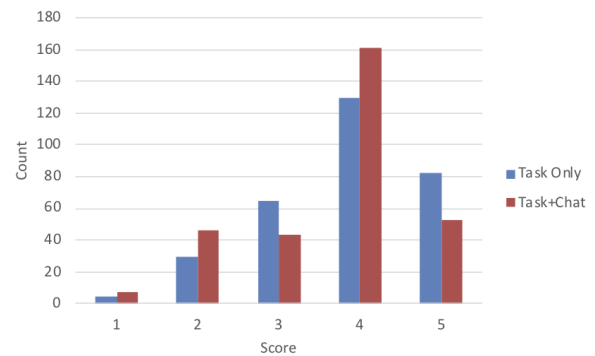


Figure 4: Distribution of Ratings Overall (Counts Summed for Questionnaire Items Across Both Experiences).

A series of Wilcoxon Signed Ranks Tests did show that participants were significantly happier to talk to the task-only system again in the future compared to the task+chat condition version following repeat use of the bots ($p = 0.027$). Figure 5 shows the distribution of scores. No other significant differences in attitude were found between the two conditions.

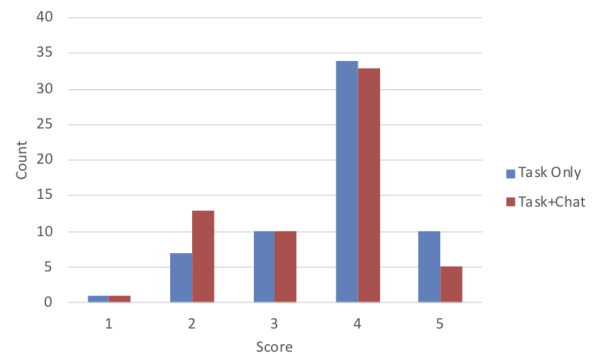


Figure 5: Distribution of Ratings for "Happy to Talk Again" Following Second Use.

5.1.2 Explicit Preference. When asked which version of the system they preferred, there was a tendency for participants to choose the task-only version (62.5%); however, this result was not statistically significant (binomial test, excluding 6.3% who expressed no preference). We thus reject hypothesis H2.

5.1.3 Conversational Efficiency. Table 1 shows the results for measures of conversational efficiency. In both versions there was considerable variation in the number of conversational turns, ranging from two (the minimum required to fulfil the tasks) to over 30 in both conditions. The mean number of turns was slightly higher in the task+chat version, but this was not significant.

Measure	task-only	task+chat
Number of turns	$M=6.92, SD=4.80$	$M=7.95, SD=4.57$
Duration (secs)	$M=94.36, SD=80.76$	$M=141.14, SD=96.35$

Table 1: Conversational Efficiency Per Condition. M denotes Mean value, SD – Standard Deviation.

Despite a similar number of turns, participant interactions with the task+chat version were significantly longer compared to the task-only version ($F(1, 24) = 10.09, p = 0.004$), by an average of 46.8s or 50% (see Table 1), indicating the turns themselves in the chat version may last longer. The maximum duration was 688 secs in the task+chat version vs 518 secs in the task-only condition (33% longer). These results indicate partial support for Hypothesis H4 and are explored further below.

Participants' second interaction was on average shorter by 2.59 turns and 50 secs compared to their first ($p = 0.001$ in both cases), perhaps indicating greater willingness to explore the system's capabilities on first experience of it.

The prevalence of interactions with a high number of turns in the task-only condition was something of a surprise. Closer analysis showed that while 40.0% of responses in the task-only condition were concerned specifically with the supplied tasks, a further 34.8% of responses involved participants asking other questions relating to the GRID building on their own initiative, either requesting more detail on the supplied tasks, asking variations on them (e.g. directions to other rooms), or enquiring about completely different aspects of the building e.g. "Are there any showers in the building?" These types of enquiry, which we have termed 'Off-task GRID', occurred in both experiment conditions.

Table 2 shows the distribution of response types of key interest across both experiment conditions. Other response types not listed included confirmations / clarifications, 'goodbye's, etc.

User response type	task-only	task+chat
Task-specific	36.6%	36.6%
Off-task GRID	33.4%	20.8%
Social	0.7%	14.3%

Table 2: Key User Response Types Per Condition.

In the task+chat version, a similar proportion of user utterances to the task-only condition were specifically task-related (36.6%). Responding to the availability of chat features, some 14.3% were on social topics (e.g. "What's your favourite Taylor Swift song?"), but these types of response were still outweighed by off-task enquiries related to the GRID building ('Off-task GRID' - 20.8%).

Analysis showed that only twelve of the 32 participants (37.5%) engaged in forms of social chat with the task+chat version. Moreover, just four people accounted for 63.5% of the chat responses. In contrast, 21 of the participants made off-task GRID-related enquiries in the task+chat condition (in the task-only it was 25).

In terms of conversational depth, once users embarked on social chat the number of turns ranged from a single user-system pair to a maximum of twelve before the participant either ended the interaction or reverted to task or GRID-based dialogue. The mean number of social turns was 3.41 ($SD = 2.67$).

Given the relatively low incidence of social chat, the analysis of interaction duration and number of turns was re-run excluding the twelve participants who had taken part in social conversation at least once. Overall, however, the pattern remained the same. Participant interactions with the task+chat version remained significantly longer at 106.3s compared to the task-only version at 77.4s ($F(1, 12) = 9.40, p = 0.010$), albeit by a smaller proportion (by 28.9s or 37.4%). Meanwhile, the mean number of turns was very similar: 6.49 in the task-only condition and 6.33 in the task+chat version. Differences in duration therefore, cannot be explained by users engaging in social conversation. Possible explanations include the longer conversational drivers employed in the task+chat condition, which consistently advertised both the GRID functionality and the social capabilities of the system.

5.1.4 Dialogue Quality. Participants were equally successful in performing the tasks in both experimental conditions, thus verifying hypothesis H3. There was no significant difference between versions, with the mean number of tasks completed 3.91 (out of a possible 4) in the task-only condition and 3.88 in the task+chat condition (t -test, $p = 0.662$).

In both experiment conditions, a proportion of enquiries involved a user intent or goal that had not been anticipated in the system (13.3% in the task-only version and 10.3% in the task+chat condition), leading to a failure to understand the user's request. In the task-only version, all of these were off-task GRID enquiries e.g. "How much is the coffee at the cafe?" In the task+chat condition 8.9% were off-task GRID utterances, with 1.4% relating to social topics e.g. questions on Alana's likes and dislikes that had not been anticipated e.g. "What book genre do you like the most?". In such cases, as with the small number of ASR errors that led to an uninterpretable input, Coherence Bot typically answered with a generic response designed to help move the conversation forward. Both conditions were affected to a similar degree by these types of response, suggesting they are unlikely to account for the difference in dialogue duration.

5.2 Qualitative Data

Amongst participants who preferred the task-only system, the most frequent reason given (by thirteen participants or 65% of this group) was that it was more direct and/or gave them the information they wanted quicker. Eight participants (40%) referred specifically to the chat features in a negative way e.g. "The first one was asking me if I wanted to talk about music movies etc. I don't want to consume that media in this form.". Six (30.0%) specified that they wouldn't want to chat on the social topics in public / the GRID building, whilst a further three (15.0%) said some topics might be appropriate in this context (e.g. news, weather) but not others (e.g. music).

Amongst those who preferred the task+chat version, six out of ten (60%) explicitly referenced the chat features e.g. *“More fun. More intelligent, could talk about more things, not just GRID-related.”*. The others felt it understood them better.

Across both preference groups, it was suggested that the task+chat version should advertise its ability to talk about other topics less often (by a total of 21.9% of participants).

6 DISCUSSION

Based on previous work, we hypothesised that the inclusion of open-domain chat in the dialogue would positively impact user attitudes towards the system. In fact, our evaluation showed that in this context it had little effect on user attitudes towards the conversational agent as measured by the questionnaire. Participants were positive towards both versions of the system, but on the whole did not rate them significantly differently. There was only one exception: participants indicated they would be significantly happier to talk to the task-only version of the system following their second experience of each bot. We therefore reject hypothesis H1.

Despite a tendency to choose the task-only version when asked (62.5%), there was also no significant preference for one version over the other, contradicting H2. The primary reason cited for preferring the task-only version was that it was more straightforward or direct, with a sizeable minority in this group (40%) specifically commenting on the chat features in a negative way e.g. *“I don’t see myself talking about music in a building like that.”* On the other hand, a minority of participants preferred the task+chat version of the agent (31.3%), mostly citing these very same chat features as their reason. Social chat therefore holds considerable appeal for some. Interestingly, however, it was suggested across both preference groups that the task+chat version should advertise its ability to talk about other topics less often (by a total of 21.9% of participants).

In terms of dialogue quality, both approaches resulted in equally high levels of task completion, supporting hypothesis H3. As predicted, offering users the chance to engage in social chat did not negatively impact their ability to achieve their goals. Also as predicted, the offer of social chat affected the conversational efficiency; significantly increasing the duration of interactions with the system. This was in part due to the consistent use of longer system outputs that encouraged users to engage in social conversation, and which as noted above, was commented on by a minority of users. It should be noted, however, that there was little evidence that these had a negative impact on user attitudes towards this version of the system as measured by the questionnaire. Overall, participants rated the two versions of the system very similarly.

The number of conversational turns was not significantly different between the two versions, explainable due to two factors. Of key interest is the relatively low take-up of the option to engage in social chat. The majority of participants (62.5%) did not engage in social conversation with the system in either of their interactions with this version. In contrast, a large proportion asked additional off-task questions relating to the GRID building – in both experiment conditions. This is an interesting result, indicating a willingness on the part of users to go beyond the prescribed task and engage freely with the system, but in a way that appears very much influenced by the context in which it is experienced.

The data thus offer only partial support for hypothesis H4, which predicted that enabling social chat would reduce conversational efficiency. Duration was significantly increased but not primarily for the reason anticipated i.e. users chatting with the agent.

Based on our results, we posit therefore that the “walk-up” public setting and workplace nature of the environment creates a markedly different use case to the in-home, and more individual and private “companion/assistant” setting which is commonly assumed for systems like Alexa, and in which the Alana system was previously evaluated [5]. The research both highlights the importance of evaluating dialogue systems in context, and sounds a note of caution on the inclusion of social chat in public task-oriented systems.

A factor in this case may have been the nature of the chat employed by the Alana system, and the way in which it was introduced. Attempts to highlight the extent of the system’s knowledge base may have led to an over-emphasis on suggesting topics that were not immediately relevant to users. Indeed, data from the experiment suggests that chat which is more specifically related to the building and its occupants may be more appropriate, at least initially. Research has shown that ‘small talk’ on neutral subjects such as the weather or events and objects in the shared physical context builds trust between interlocutors and can act as a useful prelude to other more personal kinds of talk [3], such as disclosing personal preferences and opinions of the type sought here.

Further, while the current approach of frequently advertising the system’s capabilities is a simple way to highlight the availability of chat without impeding task dialogue for those who are not interested, it is not very natural. The dialogue management decision problem of *when* it is good to chat (or to offer to chat) needs to be further investigated. It is likely that there are specific conversational locations where users will find it more appropriate to engage in social chat than at other times. This can be approached for example as an optimisation or ranking problem using deep learning methods as in [22]. Other possibilities include adopting an adaptive dialogue policy based on a social model of the user and the current status of their relationship with the agent [8].

6.1 Design considerations for Conversational AI in professional buildings

The evaluation results give rise to the following considerations when developing conversational AI in a non-private non-individual, and working setting such as a public or corporate building:

- the primary focus of the system should be on relevant information and task completion in the building itself,
- while being very desirable for some users at some times, social open-domain chat should be offered to users only occasionally and on-demand, rather than being offered equally or more often than building/task-related information,
- the public aspects of such use cases mean that users are less willing to discuss social and personal preferences (e.g. in music and movies) than in more private settings. Topics relevant to the shared physical/spatial context may be preferred.

7 CONCLUSIONS AND FUTURE WORK

In this paper we presented and evaluated a deployed conversational AI system that acts as a host of a university campus building. In an

experiment with 32 users we investigated the benefits of combining open-domain social chat with task-based dialogue; comparing our socially-enabled bot with a task-only version of the system. In fact, the results showed little evidence of benefit, with some negative effects on conversational efficiency. A minority of participants chose to engage in social conversation, instead pursuing additional enquiries relating to the building. Comments from the interviews suggested disinclination to engage in social chat may have been influenced by the system's public setting, although the experiment context with its focus on tasks may also have been a factor. Task completion was, however, unaffected by the ability to chat, with users equally able to achieve their goals with this version of the system. There was, moreover, little difference in attitudes towards the two. Participants rated both versions of the system positively and did not significantly prefer one over the other.

The experiment has highlighted areas for improvement in the coverage of enquiries related to the GRID building (both versions were equally affected) providing valuable data with which to improve the system.

Other plans for future work include making the Alana-GRID system available to users on Android phones via Google Actions. Potential benefits include a greater degree of discretion over when and where to interact with the system. Indications from the experiment reported here are that the option to interact with the system in more private circumstances may lead to a greater interest in, and uptake of, its conversational abilities.

ACKNOWLEDGMENTS

This work has been partially funded by the EU H2020 program under grant agreement no. 871245 (SPRING)³ and by an EPSRC Impact Acceleration Grant.

REFERENCES

- [1] Bayan Abushawar and Eric Atwell. 2015. ALICE chatbot: Trials and outputs. *Computación y Sistemas* 19 (12 2015). <https://doi.org/10.13053/cys-19-4-2326>
- [2] Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. *SIGCHI* (01 2001). <https://doi.org/10.1145/365024.365304>
- [3] Timothy Bickmore, Laura Vardoulakis, and Daniel Schulman. 2013. Tinker: A relational agent museum guide. *Autonomous Agents and Multi-Agent Systems* 27 (12 2013). <https://doi.org/10.1007/s10458-012-9216-7>
- [4] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explorations* 19 (2017), 25–35.
- [5] Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalymov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings* (2018).
- [6] Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup, Ioannis Papaioannou, Andrea Vanzo, Jean-Marc Odobez, Olivier Canévet, Yuanzhouhan Cao, Weipeng He, Angel Martínez-González, Petr Motlíček, Rémy Siegfried, Rachid Alami, Kathleen Belhassein, Guilhem Buisan, Aurélie Clodic, Amandine Mayima, Yoan Sallami, Guillaume Sarthou, Phani-Teja Singamaneni, Jules Waldhart, Alexandre Mazel, Maxime Caniot, Marketta Niemelä, Päivi Heikkilä, Hanna Lammi, and Antti Tammela. 2019. MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces.
- [7] A. L. Gorin, G. Riccardi, and J. H. Wright. 1997. How May I Help You? *Speech Commun.* 23, 1–2 (Oct. 1997), 113–127. [https://doi.org/10.1016/S0167-6393\(97\)00040-X](https://doi.org/10.1016/S0167-6393(97)00040-X)
- [8] Alankar Jain, Florian Pecune, Yoichi Matsuyama, and Justine Cassell. 2018. A User Simulator Architecture for Socially-Aware Conversational Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) (IVA '18). Association for Computing Machinery, New York, NY, USA, 133–140. <https://doi.org/10.1145/3267851.3267916>
- [9] Saba Khashe, Gale Lucas, Burcin Becerik-Gerber, and Jonathan Gratch. 2019. Establishing Social Dialog between Buildings and Their Users. *International Journal of Human-Computer Interaction* 35, 17 (2019), 1545–1556. <https://doi.org/10.1080/10447318.2018.1555346>
- [10] Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal, Dilek Hakkani-Tur, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize. In *Alexa Prize Proceedings*. 1–28. arXiv:1812.10757 <http://arxiv.org/abs/1812.10757>
- [11] Oliver Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. *8th Annual Conference of the International Speech Communication Association, Interspeech 2007* 3 (2007), 1761–1764.
- [12] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* 140 (1932).
- [13] Irene Lopatovska and Heyrling Oropeza. 2018. User interactions with “Alexa” in public academic space. *Proceedings of the Association for Information Science and Technology* 55, 1 (2018), 309–318. <https://doi.org/10.1002/pra2.2018.14505501034> arXiv:<https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2018.14505501034>
- [14] Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalymov, Xu Xinnuo, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. 2017. An Ensemble Model with Ranking for Social Dialogue. *NIPS 2017 Conversational AI Workshop*; Conference date: 08-12-2017 Through 08-12-2017.
- [15] Ioannis Papaioannou, Amanda Cercas Curry, Jose L. Part, Igor Shalymov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, et al. 2017. Alana: Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback. *Alexa Prize Proceedings* (2017).
- [16] Ioannis Papaioannou, Christian Dondrup, Jekaterina Novikova, and Oliver Lemon. 2017. Hybrid chat and task dialogue for more engaging HRI using reinforcement learning. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*. IEEE, 593–598.
- [17] Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field Trial Analysis of Socially Aware Robot Assistant. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) (AAMAS '18). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1241–1249.
- [18] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. “Alexa is my new BFF”: Social Roles, User Satisfaction, and Personification of the Amazon Echo. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17* (2017), 2853–2859. <https://doi.org/10.1145/3027063.3053246>
- [19] Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi. 2005. Let's Go Public! Taking a spoken dialog system to the real world. *9th European Conference on Speech Communication and Technology* (2005), 885–888.
- [20] Emanuel A. Schegloff. 1968. Sequencing in Conversational Openings. *American Anthropologist* 70, 6 (1968), 1075–1095.
- [21] Alex Sciuto, Armita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. “Hey Alexa, What's Up?": Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18* (2018), 857–868. <https://doi.org/10.1145/3196709.3196772>
- [22] Igor Shalymov, Ondřej Dušek, and Oliver Lemon. 2018. Neural Response Ranking for Social Conversation: A Data-Efficient Approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. Association for Computational Linguistics, Brussels, Belgium.
- [23] Weronika Sieińska, Nancie Gunson, Christopher Walsh, Christian Dondrup, and Oliver Lemon. 2020. Conversational Agents for Intelligent Buildings. *Proceedings of the 2020 SIGDial Conference* (2020).
- [24] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the 8th Conference on European Chapter of the Association of Computational Linguistics, ACL '97*. 271–280.
- [25] Steve Young, Milica Gašić, Blaise Thomson, and Jason Williams. 2012. POMDP-based Statistical Spoken Dialogue Systems: a Review. In *Proceedings of the IEEE* (proceedings of the ieee ed.), 1–20. <https://www.microsoft.com/en-us/research/publication/pomdp-based-statistical-spoken-dialogue-systems-a-review/> DOI 10.1109/JPROC.2012.2225812.
- [26] Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. 2017. Learning conversational systems that interleave task and non-task content. *IJCAI International Joint Conference on Artificial Intelligence* (2017), 4214–4220. <https://doi.org/10.24963/ijcai.2017/589> arXiv:1703.00099
- [27] Zhou Yu, Dan Bohus, and Eric Horvitz. 2015. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. *SIGDIAL 2015 September* (2015), 402–406.

³See <http://spring-h2020.eu/>